Nombre del proyecto
Automatización de facturas vehiculares

Profesor	Mgt. Gladys María Villegas Rugel
Materia	PROCESAMIENTO DE LENGUAJE NATURAL
Alumnos	Andrea Fernanda Morán Vargas Pedro José Vidal Orús
Fecha	19 septiembre 2025

Resumen del Proyecto

Automatización de Facturas Vehiculares con IA

1. Contexto del Problema

Dominio: Finanzas – Créditos vehiculares.
Reto actual: Procesar facturas en PDF o imágenes con gran variabilidad de formatos y calidad.
Limitaciones: Procesos manuales demandantes, alta tasa de error humano, fallas en RPA o reglas fijas.
Riesgos: Reprocesos, demoras en desembolsos, posibles brechas de cumplimiento y fraude.

2. Solución Propuesta

Arquitectura de IA híbrida (visión + lenguaje):

- OCR avanzado: Lectura de PDFs, imágenes y escaneos.
- LLM de extracción: Normalización de datos en JSON.
- Reglas de negocio automáticas:
  - Totales coherentes con ítems.
  - Validación de IVA, VIN (17 caracteres), moneda y fecha.
  - Mecanismo de confianza: Campos con baja probabilidad se derivan a un validador humano vía interfaz.
- Retroalimentación continua: Correcciones enriquecen el dataset para mejorar desempeño (fine tuning).

3. Beneficios

- Reducción significativa de carga operativa manual.
- Mayor precisión y control frente a errores y fraudes.
- Escalabilidad ante cambios de formato en facturas.
- Cumplimiento regulatorio más confiable.

4. Viabilidad

Datos requeridos: 10–40 facturas vehiculares diversas para entrenamiento inicial.
Tiempo estimado: 6 semanas (flujo, integración y pruebas).
Recurso clave: LLM + OCR + validación asistida.

Actividad	Fecha
Automatización de facturas vehiculares	19 septiembre 2025

Tabla de Contenido

1. MAPA DEL ESTADO DEL ARTE **3**
Avances Recientes en Extracción de Documentos (2018–2023) 3
Enfoques Principales Existentes y Comparativa 5
Gaps Identificados y Posicionamiento de la Propuesta 7
2. ANÁLISIS DE DATASETS DISPONIBLES **8**
Dataset 1: SROIE – Recibos Escaneados (ICDAR 2019) 9
Dataset 2: IDSEM – Facturas de Electricidad en España (2022) 9
Dataset 3: DocILE – Documentos Empresariales Multi-formato (ICDAR 2023) 11
3. DEFINICIÓN DE MÉTRICAS DE ÉXITO **14**
Métricas técnicas (desempeño del modelo e infraestructura) 14
Métricas de negocio (Valor para el banco) 14
Métricas de usabilidad (Experiencia de los usuarios) 15
4. ANÁLISIS DE STAKEHOLDERS **15**
Stakeholders primarios (usuarios finales o beneficiarios del sistema) 16
Stakeholders secundarios (Entorno externo, reguladores y actores indirectos) 16
Stakeholders claves (Sponsors, financistas, expertos en negocio o implementadores con alto poder de decisión) 16
Matriz de Interés e Influencia – Automatización de Facturas Vehiculares 17
5. DOCUMENTO DE ALCANCE DEL PROYECTO **19**
OBJETIVO GENERAL 19
OBJETIVOS ESPECÍFICOS 19
ALCANCE INCLUIDO 19
ALCANCE EXCLUIDO 23
CRITERIOS DE ACEPTACIÓN 24
6. CRONOGRAMA CON METODOLOGÍA ÁGIL **27**
Metodología Scrum (roles y ceremonias) 29
7. PLAN DE RECURSOS **29**
Recursos Humanos 29
Recursos Técnicos 31
Hardware 31
Software 32
Datos 33
Consideraciones clave en los recursos técnicos 34
Recursos Financieros 34
8. HITOS Y ENTREGABLES **36**
9. REFERENCIAS **37**

1. MAPA DEL ESTADO DEL ARTE

Visión general:

La extracción automatizada de información de documentos financieros (como facturas o recibos) ha avanzado significativamente en los últimos años. A continuación, se presenta una revisión de la literatura reciente enfocada en la extracción de datos clave desde imágenes o PDFs de facturas, destacando enfoques principales, sus fortalezas/debilidades, y el hueco identificado que motiva la propuesta actual.

Avances Recientes en Extracción de Documentos (2018–2023)

- **2018:** Surgen modelos basados en visión que capturan la disposición 2D del texto. Un hito fue Chargrid (EMNLP 2018), que representa el documento como una cuadrícula de caracteres y usa una red completamente convolucional para segmentar y extraer campos. Este enfoque logró extraer información de facturas superando claramente métodos secuenciales tradicionales basados solo en texto plano, demostrando la importancia de la información espacial en documentos (Katti et al., 2018).

- **2019:** Se reconoce formalmente el desafío de Key Information Extraction (KIE) en documentos. En ICDAR 2019 se lanza la competencia SROIE con 1000 recibos escaneados para evaluar OCR y extracción de campos clave (Huang et al., 2021). Además, se publican datasets especializados como CORD (Consolidated Receipt Dataset), con más de 11 000 recibos anotados para tareas de post-OCR parsing (Park et al., 2019). Esta época marca el interés por combinar OCR con NLP para entender documentos semiestructurados.

- **2020:** Introducción de modelos pre-entrenados multimodales específicos para documentos. Microsoft lanza LayoutLM (ACL 2020), primer modelo que incorpora embeddings 2D de posición junto con el texto, aprendiendo conjuntamente el contenido textual y su ubicación (Bevin et al., 2025). Esto representó un gran avance en Document Understanding, logrando mejoras notables en tareas de clasificación y extracción de formularios.

- **2021:** Avances en modelos multimodales más sofisticados. LayoutLMv2 (ACL 2021) añade características visuales de la imagen para hacer el modelo verdaderamente multimodal (Bevin et al., 2025). Amazon presenta DocFormer (ICCV 2021), introduciendo una arquitectura transformer de múltiples modalidades (texto, visión, posición) para entender documentos de forma integral. Estos modelos establecen nuevos state of the art en benchmarks como FUNSD (formulario semiestructurado) y CORD, demostrando la eficacia de fusionar visión y lenguaje.

- **2022:** Surgen enfoques generativos y end-to-end. LayoutLMv3 (2022) simplifica la arquitectura y unifica objetivos de texto-imagen, logrando el nuevo estado del arte en múltiples tareas de Document AI (Bevin et al., 2025). Al mismo tiempo, se propone Donut (Document Understanding Transformer, ECCV 2022), un modelo encoder-decoder que elimina la dependencia de OCR tradicional, aprendiendo a leer el documento como imagen y generar directamente un JSON estructurado de salida (Bevin et al., 2025). Donut demostró desempeño competitivo en extracción de información con mayor velocidad al evitar la etapa OCR (Kim et al., 2021). También aparecen técnicas híbridas de generación de datos sintéticos para paliar la falta de datos anotados; por ejemplo, generar facturas sintéticas preservando el layout original y reemplazando campos mediante LLMs (Bevin et al., 2025).

- **2023:** Adopción de Large Language Models (LLMs) generalistas en el dominio de documentos. Con la aparición de GPT-3/4 y similares, se empiezan a usar LLMs para interpretar resultados OCR o incluso para leer documentos directamente. Empresas como Uber reportan la sustitución de RPA por motores generativos: en 2023, Uber integró GPT-4 en su pipeline de facturas, logrando reducir el tiempo de procesamiento en 70% y duplicar la productividad, con ahorros de 25–30% (Lin, 2025). Asimismo, se publica DocILE (ICDAR 2023), el mayor benchmark a la fecha con 6 680 documentos empresariales reales y 100 000 sintéticos para tareas de localización y extracción de campos clave en facturas, órdenes de compra, etc. (Šimsa et al., 2023). Esto evidencia la tendencia hacia evaluaciones a gran escala y la importancia de enfoques que aprovechen datos masivos sintéticos y reales. En resumen, 2023 consolida la idea de combinar OCR avanzado con la capacidad de razonamiento contextual de los LLMs, permitiendo una comprensión semántica más allá de plantillas rígidas.

Enfoques Principales Existentes y Comparativa

A partir de la literatura revisada, se identifican cinco enfoques principales para la extracción automatizada de información en facturas/documentos

Enfoque	Descripción	Fortalezas	Debilidades
Basado en Plantillas y RPA	Extracción mediante reglas predefinidas por formato, como los RPA tradicional con if/then.	Simplicidad inicial. Útil si todos los documentos siguen un formato fijo.	Frágil ante variaciones. Un cambio en el layout quiebra el script Alto costo de mantenimiento por cada nuevo proveedor/format Escalabilidad muy limitada (Lin, 2025)
OCR + Reglas NLP Secuencial	Usar OCR para obtener texto y luego algoritmos NLP secuenciales (regex, CRF, etc.) para etiquetar entidades en el texto lineal.	Aprovecha texto legible. Más flexible que plantillas puras. Puede aplicar validaciones simples como regex.	Ignora la estructura 2D del documento, el orden de lectura puede ser incoherente. (Katti et al., 2018) Dificultad con campos dispersos en diferentes zonas. Requiere bastante ajuste manual de reglas, propenso a errores si OCR falla en algún punto.
Modelos Visión+Texto (CNN/Graph)	Modelos entrenados end-to-end que consideran la posición de palabras, o grafos	Capturan relaciones espaciales de los campos, mejorando la precisión sobre métodos secuenciales.	Necesitan datasets etiquetados grandes para entrenar. Arquitectura compleja, entrenar CNN/GCN en

	donde nodos son palabras y aristas su proximidad.	No dependen de plantillas explícitas, pueden generalizar a layouts no vistos.	2D requiere muchos recursos.
Transformers Multimodales Pre-entrenados	Modelos tipo LayoutLM, DocFormer, FormNet, etc., pre-entrenados en grandes corporas de documentos.	Estado del arte en muchos benchmarks gracias al pre-entrenamiento. Entienden contexto textual y estructura simultáneamente.	Tamaño grande e inferencia más lenta, requiere GPUs. Requiere fine-tuning con datos anotados del dominio para máximo desempeño. Muchos modelos base entrenados principalmente en inglés.
Enfoques Generativos con LLM	Uso de Large Language Models para comprender y extraer información con ayuda de modelos OCR y LLM Vision.	Adaptabilidad. Los LLM aprenden en contexto, pueden manejar formatos nunca vistos sin reprogramación. Reducción drástica de reglas fijas. El modelo entiende semántica (ej. reconoce un VIN aunque la etiqueta cambie). Permite incorporar reglas de negocio en el prompt (ej. validar sumas de forma dinámica).	Posible “alucinación” o errores no trazables del LLM (inventar campos si OCR fue ambiguo). Optimizar prompts o fine-tuning puede requerir iteraciones para garantizar JSON válido, etc.

Los enfoques tradicionales (plantillas, OCR+regex) tienden a fallar ante la variabilidad y calidad heterogénea de facturas reales (Lin, 2025). En cambio, los modelos con comprensión layout+texto (p.ej. LayoutLM) han mostrado F1 >90% en campos clave en benchmarks públicos (Bevin et al., 2025), y soluciones industriales con LLM reportan mejoras sustanciales: Uber alcanzó ~90% de exactitud promedio y redujo un 90% el trabajo manual en su proceso de facturas al combinar OCR con GPT-4 (Lin, 2025). Esto confirma que los

LLM con visión pueden superar las limitaciones de OCR puro, al inferir significado y corregir errores contextualmente (ej., deducir un número ilegible por contexto) (Retnan, 2025).

Gaps Identificados y Posicionamiento de la Propuesta

A pesar de los avances, existen limitaciones clave en el estado actual que abren espacio para nuestra propuesta:

Alta variabilidad no resuelta completamente: Las facturas vehiculares en particular presentan múltiples formatos (cada concesionaria tiene su diseño con logos, distintos campos de descuentos, etc.) y a menudo vienen en calidad baja (escaneos, fotos de clientes). Los métodos basados en plantillas o reglas fijas fallan rotundamente en estos escenarios (Lin, 2025). Incluso las soluciones actuales requieren un esfuerzo considerable de fine-tuning o creación de datos para cada nuevo formato, lo cual es un problema abierto (Sánchez, Salgado, García & Monzón, 2022).

Campos especializados (dominio local): Campos como el VIN (número de chasis de 17 caracteres) o el RUC no aparecen explícitamente en los datasets públicos globales revisados. Esto significa que los modelos pre-entrenados generales podrían no reconocer inmediatamente estos campos o sus validaciones. Hay un vacío de datos público en el subdominio de facturas de vehículos en español, lo que obliga a crear muestras propias o adaptar modelos con datos de otros tipos de facturas.

Combinación de visión y semántica: Muchos enfoques se centran en una sola técnica. Nuestra propuesta se posiciona en combinar la robustez del OCR especializado de Azure (para lidiar con texto en imágenes de baja calidad, sellos, firmas, etc.) con el razonamiento semántico de un LLM ajustado al dominio vehicular. Creemos que esta sinergia aborda el gap donde ni el OCR por sí solo (que da texto pero no entiende contextos) ni un LLM solo (que sin ver bien puede alucinar) son suficientes. En efecto, Uber y otros casos han mostrado que esta combinación reduce excepciones y mantenimiento (Lin, 2025).

Necesidad de aprendizaje continuo: Dado que los formatos evolucionan, la solución debe aprender de nuevos ejemplos. Los enfoques rígidos no lo permiten fácilmente. Nuestra propuesta incluye un bucle de retroalimentación (Human-in-the-Loop) para que cada corrección humana alimente un ajuste fino futuro del modelo de lenguaje. Esto responde a la

limitación actual de LLMs que, si bien aprenden contexto, requieren actualización para nuevos estilos o vocabularios específicos.

La solución propuesta se apoya en el estado del arte pero apunta a un caso innovador en Ecuador: automatizar la verificación de facturas vehiculares usando IA.

A diferencia de soluciones generales, se adaptará a documentos locales (concesionarias como Autolasa, Ambacar, Maresa) mediante un fine-tuning específico del modelo de lenguaje con ejemplos reales del mercado ecuatoriano. Esto llenará el vacío de no tener modelos especializados en español para este tipo de factura. En suma, la propuesta se alinea con la tendencia de OCR + LLM que ha demostrado ser superior en flexibilidad y precisión (Lin, 2025), y la lleva a un contexto novedoso, garantizando también controles de validación (sumas coherentes, formato de VIN, cálculo de IVA) para asegurar un resultado confiable.

Se espera así mejorar la precisión en campos críticos por encima del 95% (métrica comparable a casos reportados) y reducir el tiempo de procesamiento manual en ≥90%, cerrando brechas de cumplimiento y fraude identificadas en el proceso actual.

2. ANÁLISIS DE DATASETS DISPONIBLES

Dataset 1: SROIE – Recibos Escaneados (ICDAR 2019)

Descripción: SROIE (Scanned Receipts OCR and Information Extraction) es un dataset publicado para la competencia ICDAR 2019 de lectura robusta. Contiene ~1000 imágenes de recibos de compras reales (formato JPEG) con sus textos anotados. En la tarea de extracción (Task 3) se debían obtener campos clave como nombre de tienda, dirección, fecha y total de cada recibo, almacenándolos en JSON.

Procedencia: Los datos fueron recopilados por los organizadores de ICDAR en 2019; provienen de recibos en idioma inglés (ej. comercios asiáticos con texto en inglés). Es de acceso público para investigación (disponible en Kaggle y repositorios académicos sin costo).

Calidad: Las imágenes incluyen variabilidad moderada: diferentes layouts de recibos de tiendas, algunos con impresiones térmicas de baja calidad, leves rotaciones o borrones típicos. El OCR es desafiante pero factible; el dataset provee anotaciones textuales completas y los valores esperados de

campos. Al ser escaneos del mundo real, refleja bien ruidos y complejidades (p.ej., fuentes poco comunes, sellos).

Limitaciones/Sesgos: Es un dataset relativamente pequeño (solo 1000 ejemplos, de los cuales ~600 entreno). Además, está limitado al dominio de retail: los campos son propios de recibos de tienda (no incluyen, por ejemplo, identificadores de cliente ni campos vehiculares). El idioma inglés y contextos culturales diferentes implican que los formatos difieren de las facturas ecuatorianas. No contiene ningún VIN, RUC u otros campos especializados que necesitamos.

Idoneidad para nuestro proyecto: SROIE sirve como punto de referencia inicial para probar algoritmos de OCR+extracción, y sus campos de total, fecha etc. son análogos a los de una factura vehicular. Sin embargo, por idioma y contexto, no es directamente aplicable para entrenar un modelo en español o con nuestros campos específicos. Podría utilizarse para pre-entrenar o evaluar la capacidad base del modelo (p.ej., ver si el modelo detecta totales en documentos sencillos), pero requeriría transfer learning significativo.

Accesibilidad: Alta – Disponible públicamente. Se puede descargar desde Kaggle u otros repositorios sin restricciones conocidas. No contiene PII sensible (son datos ficticios o comerciales genéricos), por lo que su uso no tiene trabas legales.

Dataset 2: IDSEM – Facturas de Electricidad en España (2022)

Descripción: IDSEM (Invoice Dataset of Spanish Electricity Market) es una base de datos publicada en Scientific Data (Nature) en 2022. Consiste en 75 000 facturas de electricidad simuladas en formato PDF, con sus correspondientes etiquetas estructuradas en archivos JSON. Cada factura incluye ~86 campos distintos que abarcan datos del cliente, empresa eléctrica, consumo, importes, impuestos, fechas, etc.. Los documentos fueron generados siguiendo formatos reales del mercado eléctrico español.

Procedencia: Dada la dificultad de obtener facturas reales (por privacidad), los autores emplearon un proceso de simulación basado en regulaciones y estadísticas reales para generar las facturas. Partieron de plantillas de varias comercializadoras españolas y poblaron datos sintéticos realistas. El resultado es un dataset balanceado conforme a casos reales pero con información ficticia. Está publicado con acceso abierto (Open Access) y pensado justamente para entrenar algoritmos de extracción de información.

Calidad: Aunque las facturas son sintéticas, mantienen alta fidelidad a las reales: incluyen logotipos, distintas distribuciones de secciones, varios estilos de tabla de consumo, etc. Al ser PDFs, la calidad

visual es excelente (texto digital claramente legible). La variabilidad proviene de usar múltiples compañías y escenarios de consumo, pero dado que el sector está regulado, la estructura base es similar en todas (por ejemplo, siempre hay un apartado de detalles de consumo, uno de impuestos, totales, etc.). Las etiquetas JSON son completas y precisas (sin errores humanos, por ser generadas automáticamente). Esto es ideal para entrenamiento supervisado.

Limitaciones/Sesgos: Por un lado, no son documentos escaneados: no presenta ruidos de fotografía o escaneo, por lo que un modelo entrenado solo aquí podría rendir menos en imágenes reales con ruido. Además, el dominio es facturas de electricidad, con campos específicos (p. ej., número de contrato, consumo kWh, código CUPS) que no se requieren en una factura vehicular, mientras faltan otros que sí necesitamos (VIN, datos del vehículo). Aun así, muchos campos son análogos: cliente (comprador), empresa emisora (vendedor), subtotal, impuestos (IVA), total, fechas. El idioma es español de España (terminología ligeramente distinta a Ecuador en algunos casos) y la moneda euros. Puede haber un sesgo de formato relativamente homogéneo debido a la regulación común – menos diversidad que en facturas de distintas industrias.

Idoneidad: IDSEM es altamente relevante por estar en español y proveer un volumen enorme de ejemplos. Un modelo podría pre-entrenarse o afinarse con estas 75k facturas para aprender a reconocer campos en documentos en español con estructura de factura. Esperamos que aprenda a identificar etiquetas como “Subtotal”, “IVA 12%”, “Total a pagar”, etc., que son similares en nuestras facturas. Sin embargo, habría que exponer al modelo a ejemplos de menor calidad (escaneos) posteriormente, ya que aquí todo es digital. Y se tendría que transferir el aprendizaje a las facturas vehiculares incorporando unos cuantos ejemplos reales de ese dominio para añadir el campo VIN y ajustar diferencias (por ejemplo, en una factura de auto podría haber campo de placa, que no existe en una eléctrica).

Accesibilidad: Media: Si bien es open access, el dataset puede requerir contactar a los autores o descargar de un repositorio asociado a Nature. Al ser 75k PDFs (~ gigabytes de datos), manejarlo completo implica recursos significativos. No obstante, para investigación es obtenible gratuitamente. La licencia debería permitir uso no comercial libremente (Scientific Data suele publicar bajo Creative Commons). Un aspecto a considerar es el costo temporal de procesar tantos PDFs y JSON; quizás se deba muestrear o utilizar parcialmente debido al límite de tiempo del proyecto.

Dataset 3: DocILE – Documentos Empresariales Multi-formato (ICDAR 2023)

Descripción: DocILE (Document Information Localization and Extraction) es un nuevo benchmark a gran escala para documentos de negocio semiestructurados. Reúne ≈106 680 documentos de tipo factura, órdenes de compra, etc., de los cuales 6 680 son reales anotados manualmente y ~100 000 son

sintéticos, además de casi 1 millón de documentos no anotados para pre-entrenamiento. Las anotaciones cubren dos tareas: KILE (Key Information Localization & Extraction), es decir, detectar y extraer campos clave como número de factura, fecha, total, nombres de las partes; y LIR (Line Item Recognition), es decir, extraer las líneas de detalle (productos, cantidades, precios unitarios).

Procedencia: Fue compilado por Rossum AI y colaboradores, a partir de documentos de empresas y también generando sintéticos para ampliar. Incluye datos de múltiples países e idiomas (principalmente facturas en inglés, pero potencialmente algunos documentos en otros idiomas europeos; se diseñó para evaluar métodos independientes del idioma). Es un dataset de investigación; para acceder se debe completar un formulario de solicitud (debido a posibles datos sensibles en algunos documentos reales), pero está destinado a uso académico gratuito.

Calidad: Es el dataset más variado: contiene documentos de distintos proveedores, industrias y formatos, cubriendo muchos escenarios del mundo real. Hay documentos escaneados (con ruido, escrituras a mano, sellos) y otros digitales. Los campos anotados permiten probar exhaustivamente un extractor. La inclusión de 100k sintéticos ayuda a que haya volumen, aunque los sintéticos pueden ser menos variados que los reales. En general, la calidad de anotación es alta (revisada en un contexto de competencia). Para nuestro caso, aporta ejemplos de facturas generales muy útiles (los campos que buscamos como total, fecha, etc. están definitivamente incluidos).

Limitaciones: Por su naturaleza amplia, no está focalizado en facturas vehiculares ni en español. La mayoría de datos está en inglés y aunque abarca “invoices”, no distingue tipo de factura por sector; una factura de coche es tratada como cualquier factura. No esperamos que contenga VIN o RUC, a menos que incidentalmente alguna factura real sea de un concesionario (difícil de saber). Además, el tamaño masivo dificulta usarlo completo en un proyecto corto – entrenar desde cero con DocILE está fuera de alcance por tiempo. Otro factor es la accesibilidad controlada: hay que solicitarlo, lo cual puede tomar tiempo y posiblemente filtrado de datos (quizá excluyan información que pueda considerarse privada).

Idoneidad: DocILE es excelente para evaluar nuestro método frente a los mejores, dado que es un benchmark público. Podríamos, por ejemplo, probar nuestro modelo afinado en español para ver cómo rinde en algunas facturas en inglés de DocILE (solo como comparación). Sin embargo, para entrenamiento principal no es práctico. Una posibilidad sería aprovechar los 100k documentos sintéticos de DocILE, que podrían incluir plantillas de facturas variadas, pero nuevamente el obstáculo idiomático está. En resumen, es más útil como conjunto de validación global o fuente de inspiración en cuanto a qué campos extraer, que como fuente directa de entrenamiento.

Accesibilidad: Baja/Media: requiere solicitar acceso (lo cual concede un enlace de descarga). No hay costo monetario, pero hay restricciones de uso (investigación, no uso comercial, etc.). Dado el tiempo limitado, es posible que no lleguemos a incorporarlo de manera significativa, por lo que se consideraría sólo si los demás datasets no son suficientes.

Dataset	Tipo de documentos	Tamaño	Idioma	Ventajas	Desventajas
SROIE	Recibos de compra	1000	Inglés	Datos reales con variabilidad moderada Campos clave anotados fáciles de mapear.	Muy pequeño. Contexto de retail, no vehicular. En otro idioma.
IDSEM	Facturas de luz	75000	Español	Muy amplio en número Idioma español con terminología de facturación	Sintético (menos variabilidad visual). Dominio eléctrico específico.
DocILE	Documentos empresariales variados	6 680 reales 100 000 sintéticos	Multilingüe	Dataset muy diverso en layouts Ideal para evaluar generalización Incluye tanto campos clave como ítems de tabla, con casos de imágenes ruidosas.	Acceso limitado Tamaño masivo difícil de usar completo Mayoría de documentos no en español ni del dominio vehicular Datos reales con posibles restricciones.

El dataset IDSEM se perfila como la opción principal para iniciar el entrenamiento. Su valor radica en la compatibilidad lingüística (facturas en español) y en la amplitud de sus 86 campos, que permiten entrenar un modelo inicial con buena capacidad de extracción en elementos como totales, impuestos, identificación del cliente y estructura típica de facturas

Aunque no incluye campos específicos del dominio vehicular (como VIN o chasis), sí aporta la base estadística y la familiaridad con layouts de facturas que resultan críticas para el pre-entrenamiento.

No obstante, depender únicamente de IDSEM conlleva riesgos: su origen en un sector distinto, el posible peso computacional para procesarlo en su totalidad, y la ausencia de campos automotrices clave limitan su cobertura para nuestro caso. Por ello, aunque IDSEM sea el dataset principal recomendado, es igualmente aconsejable avanzar en paralelo con el Plan B.

Plan B - Uso de ejemplos reales y templates (Scribd) para entrenamiento y testing

En Scribd (2025) se han localizado facturas reales emitidas por concesionarias concretas (por ejemplo, Autoniza, Automotores Continental S.A., Induwagen S.A., Autos Populares, Kiauto, entre otras) que contienen los campos característicos del dominio (RUC emisor/comprador, modelo, placa, VIN/chasis, subtotales, IVA y totales). Estas facturas de concesionarias reales son especialmente útiles porque reflejan variaciones auténticas en diseño y contenido que los modelos deben aprender y validar.

Análisis de distribución y extracción de layouts: Se va a extraer con OCR estos documentos reales para entender dónde y cómo aparecen los campos según cada concesionaria (variaciones de ubicación, etiquetas y formatos).

Entrenamiento y evaluación realista: Utilizar las facturas reales de estas concesionarias como parte del conjunto de entrenamiento y, muy importante, reservar una porción inalterada como test set para medir desempeño en casos verdaderos.

Complemento con generación sintética: A partir de los templates derivados de concesionarias reales, crear facturas sintéticas coherentes (variando nombres, VIN, montos) para ampliar el dataset sin depender del rotulado manual.

3. DEFINICIÓN DE MÉTRICAS DE ÉXITO

Métricas técnicas (desempeño del modelo e infraestructura)

KPI	Definición	Umbral Objetivo	Cuándo utilizar	Cálculo
F1-Score	Media armónica precision/recall	≥ 0.95	Clases desbalanceadas	2 * (Precisión * Recall) / (Precisión + Recall)
Tasa error OCR	Tasa de error OCR (Character Error Rate – CER) en campos estructurados y críticos (VIN, RUC)	<10%	Minimizar los falsos positivos	FP / (FP + VN)
Tasa intervención humana	Facturas que requieren intervención manual	<20%	Para errores de OCR o mal uso del LLM	(Facturas manuales / Total de facturas procesadas) * 100
Latencia (Response Time)	Tiempo de procesamiento por factura	Entre 5 y 30 segundos	Siempre en Tiempo real	Hora salida JSON – Hora recepción factura

Métricas de negocio (Valor para el banco)

KPI	Definición	Umbral Objetivo	Cálculo	Benchmark
Eficiencia automatización	Reducción de tiempo comparado frente al registro manual	90%	Promedio tiempo antes vs. ahora	≥ 85-95% reducción del tiempo vs. proceso manual (de horas a segundos)
Eficiencia operativa	Ahorro de costos operativos	70% de mejora	Horas-hombre liberadas / Costo de validación	Reducción de costos operativos del 40-70%
Índice de rechazo	Porcentaje de facturas rechazadas automáticamente por inconsistencias (prevención de riesgo).	≤ 5% de facturas con anomalías detectadas	(Facturas rechazadas por IA / Total facturas procesadas) × 100	Detección documental: 3–7% de anomalías relevantes detectadas automáticamente
Tasa errores humanos	Porcentaje en errores detectados en auditorías internas sobre facturas procesadas.	≥30%	(# Errores antes–Errores después)/Errores antes × 100	Reducciones del 25–40% en auditorías iniciales.

Métricas de usabilidad (Experiencia de los usuarios)

KPI	Definición	Umbral Objetivo	Cálculo	Benchmark
Tiempo promedio de corrección manual	Tiempo que toma un usuario en ajustar los campos erróneos de una factura, medido en segundos.	$\leq$ 30 seg/factura	Σ (Tiempo de corrección) / Nº de facturas corregidas	En sistemas de captura inteligente: 20–40 seg/factura.

KPI	Definición	Umbral Objetivo	Cálculo	Benchmark
Número de interacciones por factura	Cantidad de acciones que requiere un usuario para validar o corregir (clics o campos editados).	≤ 5 interacciones/factura	Σ (Clics o campos corregidos) / Nº de facturas revisadas	Benchmarks RPA/IDP: 3–6 interacciones promedio.
Satisfacción del usuario (SUS/NPS interno)	Medición de la percepción de usabilidad y satisfacción del personal que utiliza la solución.	SUS ≥ 75 / NPS ≥ 50	Promedio de resultados de encuestas SUS o NPS interno	Benchmarks de software empresarial: SUS 68 = aceptable, ≥ 80 = excelente.
Facilidad de trazabilidad	Grado de disponibilidad de logs, auditoría y trazabilidad de campos procesados por factura.	100% facturas con log completo	(Facturas con trazabilidad completa / Total facturas) × 100	Estándar en auditoría bancaria: trazabilidad obligatoria en 100% de casos.

4. ANÁLISIS DE STAKEHOLDERS

Stakeholders primarios (usuarios finales o beneficiarios del sistema)

- **Analistas operativos de crédito:** Ya no tendrán que digitar ni validar manualmente cada factura, reduciendo carga repetitiva.
- **Verificadores / Back-office de validación:** Recibirán facturas con campos parcialmente llenos y solo intervendrán en casos de baja confianza.
- **Auditores internos:** Contarán con registros trazables y menos errores que revisar.
- **Clientes solicitantes del préstamo vehicular:** Su beneficio es la obtención de un desembolso más rápido.
- **Concesionarias:** Tendrán un menor rechazo de facturas y las ventas serán más ágiles.

Stakeholders secundarios (Entorno externo, reguladores y actores indirectos)

- **Organismos reguladores** (Superintendencia de Bancos, control fiscal, tributario): tendrán acceso a información clara, completa y trazable de cada factura; verificarán los estándares de conservación y custodia digital de facturas (15 años según normativa local.
- **Proveedores de OCR/IA externos** (si se contrata tecnología de terceros): podrán retroalimentarse mejor de los errores para mejorar el modelo (ciclo de fine tuning); validarán si su tecnología genera valor frente a benchmarks internos.

- **Competidores bancarios** que presionan por eficiencia (impacto indirecto en benchmark): acceso a información secundaria que les permita comparar eficiencia (comparativos en tiempos de desembolso y error); validar que no se generen riesgos regulatorios compartidos.

Stakeholders claves *(Sponsors, financistas, expertos en negocio o implementadores con alto poder de decisión)*

- **Alta gerencia / Sponsors del proyecto** (Dirección de Operaciones, Dirección de Transformación Digital): será posible medir el retorno de inversión (ROI) y la reducción de costos operativos; visibilizarán el impacto estratégico frente a la competencia y al regulador; y podrán escalar la solución hacia otros procesos documental.
- **Área de Riesgo y Cumplimiento** (garantizan alineación normativa): podrán validar que las facturas cumplan requisitos normativos y fiscales; reducirán el riesgo de sanciones u observaciones por parte del ente regulador.
- **Área de Tecnología** (TI / Innovación, responsables de la implementación): deben demostrar que la solución es estable, escalable y compatible con la infraestructura del banco y que cumple con los estándares de ciberseguridad y privacidad definidos.
- **Área de Finanzas** (control de costos y beneficios): deberán evaluar el costo total de propiedad (licencias, desarrollo, mantenimiento), realizar el análisis de ahorro operativo por reducción de horas-hombre y evaluar los escenarios de retorno financiero en el corto y mediano plazo.
- **Consultores o expertos en negocio de crédito vehicular** (validación de reglas críticas): deben definir de las reglas críticas de validación: VIN de 17 dígitos, cálculos de IVA, coherencia de totales para que el sistema replique la lógica de negocio usada por analistas humanos; así mismo, deben promover que la automatización aumente la confianza de las áreas operativas en la información.

Matriz de Interés e Influencia – Automatización de Facturas Vehiculares

Stakeholder	Tipo	Interés	Influencia	Justificación
Analistas operativos de crédito	Primario	Alto	Medio	Son usuarios directos, reducen carga manual; poca capacidad de decisión estratégica.

Verificadores / Back-office	Primario	Alto	Medio	Beneficiarios de la automatización parcial, influyen en ajustes de usabilidad pero no en la dirección del proyecto.
Auditores internos	Primario	Medio	Medio	Su trabajo se facilita con trazabilidad y logs; influyen en definir requisitos de auditoría.
Clientes solicitantes de préstamo vehicular	Primario	Alto	Bajo	Principal beneficiario en rapidez de desembolso, pero sin poder en decisiones del sistema.
Concesionarias	Primario	Alto	Bajo	Beneficiadas con menos rechazos de facturas; no influyen en la solución técnica.
Organismos reguladores	Secundario	Medio	Alto	No usan el sistema directamente, pero definen requisitos normativos obligatorios.
Proveedores OCR/IA externos	Secundario	Medio	Medio	Pueden mejorar la solución, pero dependen del contrato con el banco.
Competidores bancarios	Secundario	Bajo	Medio	Influyen indirectamente a nivel de benchmark sectorial, no en el proyecto puntual.
Alta gerencia / Sponsors del proyecto	Clave	Alto	Alto	Deciden financiamiento, patrocinan la solución y esperan ROI visible.

Área de Riesgo y Cumplimiento	Clave	Alto	Alto	Definen alineación normativa, su validación es crítica para aprobación del sistema.
Área de Tecnología (TI / Innovación)	Clave	Alto	Alto	Ejecutores de la implementación; su criterio define estabilidad y seguridad.
Área de Finanzas	Clave	Medio	Alto	Evalúan sostenibilidad económica del proyecto, alto poder de decisión en viabilidad.
Consultores / Expertos en negocio vehicular	Clave	Alto	Medio	Aseguran que las reglas de negocio se implementen correctamente, pero no deciden financiamiento.

5. DOCUMENTO DE ALCANCE DEL PROYECTO

OBJETIVO GENERAL

Desarrollar “VeriFactura”, una solución híbrida de IA que integra visión por computador y LLM, destinada a automatizar la captura de facturas vehiculares recibidas por la banca, desde diversas fuentes y formatos, asegurando una precisión superior al 95% en datos críticos, utilizando datos reales de Scribd complementados con generación sintética limitada, para ser implementada en 6 semanas.

OBJETIVOS ESPECÍFICOS

Durante el desarrollo de “VeriFactura” se pretende:

- Alcanzar un F1-Score mínimo de 0.95 en la extracción automática de datos críticos de facturas vehiculares, que asegurar alta precisión y equilibrio en la clasificación de datos.
- Mantener la tasa de error OCR en campos críticos (como VIN y RUC) por debajo del 10%, minimizando falsos positivos garantizando la exactitud del procesamiento.

- Reducir la necesidad de intervención manual a menos del 20%, optimizando el uso de IA híbrida para minimizar errores que requieren corrección humana.
- Asegurar que el tiempo de procesamiento por factura se mantenga entre 5 y 30 segundos, garantizando una respuesta en tiempo real para soportar la operación bancaria continua.

ALCANCE INCLUIDO

Considerando el objetivo general y los objetivos específicos planteados, y de acuerdo al flujo de "VeriFactura" (recepción, filtrado,OCR, Modelo entrenado (LLM), resulatdos), se detallan las siguientes funcionalidades técnicas:

1. Módulo de Recepción y Pre-procesamiento de Documentos:

Ingreso de Facturas Multi-Fuente a la API: Facturas vehiculares en diversos formatos: PDF, JPG, PNG, XML.

Filtrado y Clasificación de Documentos: a través de un algoritmo identifica si el documento es un escaneo, fotografía, documento digital (PDF/Word), o texto plano.

Derivación al motor correspondiente: Dirigir el documento al motor de procesamiento apropiado: LLM para texto plano, OCR para imágenes/escaneos.

2. Módulo de Procesamiento con Lenguaje Natural (LLM) para Texto Plano:

Extracción de Información:

- **Reconocimiento de entidades nombradas (NER):** Identificar y extraer entidades relevantes como:
  - Nombre y RUC del emisor (concesionaria).
  - Nombre y RUC del receptor (banco/cliente).
  - Número de factura.
  - Fecha de emisión.
  - Chasis - VIN (Vehicle Identification Number).
  - Monto total.
  - Impuestos.
  - Marca
  - Tipo
  - Clase
  - Modelo
  - Color

- Código de motor
- RAMV
- Año

Validación y Normalización de Datos:

- **Validación de formatos:** Verificar que los datos extraídos cumplan con formatos esperados (ej. formato del RUC, formato de fecha).
- **Normalización de datos:** Estandarizar los datos extraídos (ej. convertir todas las fechas al mismo formato).

3. Módulo de Procesamiento OCR con Azure Vision Studio:

- **Integración con Azure Vision Studio:**
  - **Llamada a la API de OCR:** Invoca la API de Read de Azure Vision Studio.
- **Pre-procesamiento de Imágenes (antes del OCR):**
  - **Corrección de inclinación:** Corregir la inclinación de la imagen para mejorar la precisión del OCR.
  - **Eliminación de ruido:** Reducir el ruido en la imagen (manchas, sombras) para mejorar la calidad de la imagen para el OCR.
  - **Mejora de contraste:** Ajustar el contraste de la imagen para facilitar la detección del texto.
- **Extracción de Texto:**
  - **Reconocimiento de texto en tablas:** Identifica y extrae texto de tablas en la factura.
  - **Reconocimiento de texto en campos clave:** Prioriza el reconocimiento de texto en áreas específicas de la factura donde se espera encontrar información crítica.
- **Post-procesamiento del texto OCR:**
  - **Corrección de errores comunes de OCR:** Implementa algoritmos para corregir errores comunes de OCR (sustitución de "O" por "0", "l" por "1").
  - **Detección de idioma:** Identificar el idioma del texto OCR para facilitar la interpretación semántica.

4. Módulo de Interpretación Semántica (Post-OCR):

- **Asignación de Campos Clave:**
  - **Reglas y patrones:** Reglas y patrones basados en la disposición típica de las facturas vehiculares para asignar el texto extraído por el OCR a los campos clave (VIN, RUC, Monto total, Fecha).

    - **Uso de LLM para mejorar la asignación:** Se utiliza LLM para comprender el contexto del texto y asignar los campos clave con mayor precisión, especialmente en casos donde las reglas y patrones no son suficientes.
- **Validación y Normalización de Datos (Post-OCR):**
    - **Validación de formatos:** Verificar que los datos asignados a los campos clave cumplan con los formatos esperados.
    - **Normalización de datos:** Estandarizar los datos extraídos (convertir todas las fechas al mismo formato).
    - **Validación cruzada:** Comparar los datos extraídos de diferentes partes de la factura para asegurar la consistencia.
- **Mejora Continua del Modelo:**
    - **Almacenamiento de datos procesados:** Almacenar las facturas procesadas y los datos extraídos para entrenar y mejorar continuamente los modelos de IA.
    - **Revisión y corrección manual:** Permitir la revisión y corrección manual de los datos extraídos por el sistema.
    - **Retroalimentación al modelo:** Utilizar las correcciones manuales para retroalimentar y mejorar los modelos de IA.

5. Módulo de Interfaz de Usuario (UI) y Monitoreo:

- **Interfaz de Usuario (opcional, dependiendo de las necesidades del usuario final):**
    - **Visualización de facturas:** Permitir a los usuarios visualizar las facturas procesadas.
    - **Edición de datos extraídos:** Permitir a los usuarios editar los datos extraídos por el sistema.
- **Monitoreo:**
    - **Monitoreo del rendimiento del sistema:** Monitorear el rendimiento del sistema (KPIs definidos: tiempo de procesamiento por factura, tasa de error).

6. Requisitos No Funcionales:

- **Seguridad:** Implementar medidas de seguridad para proteger la información confidencial de las facturas (anonimización o encriptado).
- **Disponibilidad:** Asegurar que el sistema esté disponible para su uso en todo momento (KPI definido).
- **Rendimiento:** Optimizar el rendimiento del sistema para que pueda procesar las facturas en tiempo real.
- **Integración:** Facilitar la integración con los sistemas bancarios existentes.

Consideraciones Adicionales:

- **Generación de Datos Sintéticos:** Dada la limitación de datos reales de Scribd, la generación de datos sintéticos debe ser cuidadosamente considerada para complementar el entrenamiento del modelo. Esta generación debe incluir variaciones realistas en formatos, calidad de imagen (escaneo, fotografía, inclinación, etc.) y errores comunes de OCR para robustecer el modelo.
- **Pruebas Exhaustivas:** Realizar pruebas exhaustivas con datos reales y sintéticos para asegurar que el sistema cumple con los objetivos de precisión y rendimiento.
- **Iteración y Mejora Continua:** El desarrollo de "VeriFactura" debe ser un proceso iterativo, con pruebas y mejoras continuas basadas en los resultados.
- **Documentación:** Mantener una documentación completa del sistema, incluyendo la arquitectura, la implementación, y los procedimientos de operación.

ALCANCE EXCLUIDO

Con la intención de mantener el alcance del proyecto enfocado y realista dentro del plazo de 6 semanas no se incluirá en este proyecto:

Funcionalidades Fuera del Alcance de Facturas Vehiculares

Procesamiento de Otros Documentos Bancarios: No se incluirá el procesamiento de cheques, extractos bancarios, solicitudes de crédito, u otros documentos bancarios.

Soporte para Idiomas Adicionales: Inicialmente, “VeriFactura” no soportará idiomas distintos al español. La expansión a otros idiomas requeriría investigación y entrenamiento adicionales.

Funcionalidades de Automatización Bancaria Amplia

Aprobación Automática de Créditos: Aunque “VeriFactura” extraerá datos relevantes para la aprobación de créditos, NO tomará decisiones automáticas de aprobación o rechazo. La decisión final permanece en manos de los analistas bancarios.

Integración Directa con Sistemas Contables: “VeriFactura” no se integrará automáticamente con los sistemas contables de la banca. La exportación de datos estará disponible, pero la importación y reconciliación contable se manejará por separado.

Funcionalidades de Pre-procesamiento de Imágenes Extremas

Restauración de Imágenes Severamente Dañadas: "VeriFactura" no intentará restaurar imágenes con daños extremos (ej., borrosas, quemadas, rasgadas). El sistema se diseñará para manejar imágenes de calidad razonable.

Optimización Extrema para Velocidades Sub-Segundo

Procesamiento Inferior a 5 Segundos por Factura: Si bien el objetivo es estar entre 5 y 30 segundos, una optimización extrema para garantizar un procesamiento consistentemente por debajo de 5 segundos, a expensas de precisión u otras funcionalidades clave, no será una prioridad.

Justificación de Exclusiones:

Las exclusiones anteriores se basan en la necesidad de mantener el proyecto enfocado, realista y alcanzable dentro del plazo de 6 semanas. La inclusión de estas funcionalidades adicionales aumentaría significativamente la complejidad, el costo y el riesgo del proyecto. Además, algunas de estas funcionalidades podrían considerarse como mejoras futuras o fases posteriores del desarrollo.

CRITERIOS DE ACEPTACIÓN

Para del proyecto **"VeriFactura"** se han definido los siguientes criterios de aceptación, considerando:

- Condiciones específicas para considerarse completado
- Estándares mínimos de calidad
- Procedimientos de validación
-

N°	Criterio	Consideraciones	Estándares Mínimos	Procedimiento de Validación
1	**Precisión del modelo (F1-Score)**	El sistema extrae correctamente los campos críticos de las facturas vehiculares (VIN, RUC, Monto, Fecha, etc.) en un conjunto de validación.	F1-Score ≥ **0.95** en campos críticos.	Evaluación sobre conjunto de prueba mixto (real + sintético). Se calcula F1-Score global y por campo.

2	**Tasa de error OCR**	El sistema reconoce el texto en facturas escaneadas o fotografiadas, con baja cantidad de errores en campos clave.	Tasa de error OCR < **10%** (por campo).	Comparación automatizada entre texto OCR vs ground truth en campos como VIN y RUC.
3	**Reducción de intervención manual**	El sistema permite procesar facturas con mínima necesidad de corrección por parte de usuarios humanos.	Menos del **20%** de las facturas procesadas requieren corrección manual.	Registro de facturas corregidas manualmente durante las pruebas UAT. Cálculo de ratio de intervención.
4	**Tiempo de procesamiento por factura**	El sistema procesa cada factura en un tiempo eficiente desde su recepción hasta el retorno de datos.	Tiempo promedio entre **5 y 30 segundos** por factura.	Se ejecuta un lote de pruebas (ej. 100 facturas mixtas) midiendo el tiempo por factura desde input hasta output.
5	**Clasificación del tipo de documento**	El sistema determina automáticamente si un documento es imagen, escaneo o texto plano.	Clasificación correcta ≥ **95%**	Validación manual sobre lote de documentos diversos. Se compara tipo detectado vs tipo real.
6	**Derivación al motor adecuado (OCR o LLM)**	El sistema enruta correctamente cada documento al motor correspondiente para su procesamiento.	Derivación correcta ≥ **98%**	Validación cruzada entre tipo detectado y motor ejecutado. Verificación por logs y resultados.

7	**Asignación correcta de campos post-OCR**	El sistema asocia correctamente los textos extraídos por OCR a los campos estructurados correspondientes.	Asignación correcta ≥ **95%**	Revisión automatizada + validación humana sobre dataset OCRizado. Validación con reglas de layout y semántica.
8	**Validación y normalización de datos**	Los campos extraídos son verificados y transformados para cumplir con los formatos esperados.	Validación/normalización correcta ≥ **98%**	Pruebas automatizadas sobre reglas de formato (para RUC, formato de fecha, consistencia de montos).
9	**Corrección automática de errores OCR comunes**	El sistema implementa reglas de corrección para errores típicos como confusión entre letras y números.	Corrección efectiva en ≥ **90%** de casos detectados.	Prueba con documentos que contienen errores conocidos. Verificación de corrección en outputs.
10	**Visualización y edición de resultados (UI opcional)**	El usuario puede ver las facturas procesadas y editar campos si es necesario.	100% de los datos visibles y editables por UI.	Pruebas funcionales de UI: abrir factura, editar campo, guardar, y validar persistencia.
11	**Exportación de resultados estructurados**	Los datos extraídos pueden ser exportados en formato estructurado (JSON).	Exportación completa y sin pérdida de datos.	Validación de archivos de exportación generados. Verificación de correspondencia con datos visualizados.

12	**Documentación técnica completa**	Se entrega documentación detallada del sistema y su uso.	Documento con: arquitectura, API, flujo de datos, instalación y uso.	Revisión por parte de QA/TI. Lista de verificación de secciones clave.
13	**Retroalimentación para mejora continua**	Las correcciones manuales se almacenan y están disponibles para reentrenamiento del modelo.	Correcciones guardadas en > **95%** de los casos corregidos.	Verificación de logs y base de datos. Prueba de uso en ciclo de reentrenamiento simulado.
14	**Seguridad de la información**	El sistema protege los datos sensibles como RUC, VIN, montos y datos del cliente.	Cifrado en tránsito y almacenamiento implementado.	Pruebas de seguridad básicas: verificación de HTTPS, cifrado en base de datos, anonimización opcional.

6. CRONOGRAMA CON METODOLOGÍA ÁGIL

Sprint	**Duración**	**Sprint Goal**	**User Stories**	**Definition of Done**	**Riesgos y mitigación**
1	1 semana	Arquitectura y entorno listo para pipeline inicia: Detección PDF con/sin texto, OCR vs parse directo.	**US-A1** Detectar texto embebido en PDF (3pt) **US-A2** Gestor de ingesta y metadatos (5pt) **US-B2** Normalizador de texto (5pt)	**A1:** Precisión ≥95% en set de 100 PDF **A2:** Cola con reintentos y pruebas de carga mínimas (50 doc/min). **B2:** Reducción de errores OCR ≥10% vs baseline.	**R1:** OCR pobre en facturas mal escaneadas **M1:** Preprocesos (deskew/denoise)
2	2 semanas	Pipeline OCR → parser → LLM → validaciones	**US-B1** Integrar Azure OCR (5pt)	**B1:** Inferencia OCR en lote. latencia de	**R2:** Variabilidad de plantillas baja recall

		API v0 expuesta	**US-C1** Extracción LLM de RUC/VIN/totales (8pt) **US-C2** Validadores (VIN=17, sumas, RUC) (5pt) **US-D1** API REST v0 /extract con auth (5pt)	≤3s/página. **C1:** prompts/few-shot versionados. F1 inicial ≥0.85 en set de validación. **C2:** Reglas y mensajes de error. Exactitud en sumas ≥0.98 y validador de VIN. **D1:** Swagger listo.	**M2:** Few-shot por concesionaria frecuente + post-proceso por patrones. **R3:** Tiempos altos de respuesta **M3:** Batching, caché OCR, colas asíncronas y límites de tamaño.
3	2 semanas	Mejorar métricas (tuning) y validar con reportes reproducibles.	**US-F1** Script de evaluación (5pt) **US-E1** UI simple de carga/resultado (3pt) **US-C1b** Optimización de prompts por concesionaria (5pt)	**F1:** reporte reproducible. **E1:** subir doc, ver JSON y banderas de confianza **C1b:** ≥0.95 F1 en VIN/totales/RUC.	**R4:** No se alcanza F1 objetivo ≥0.95 en críticos **M4:** Híbrido reglas+LLM (regex, checks), más ejemplos y curación de casos difíciles. **R5:** Sesgos por concesionaria dominante. **M5:** Estratificar validación y balancear dataset.
4	1 semana	Pruebas integrales, documentación y demo final lista	**US-QA1** Testing integral (funcional/carga) (5pt) **US-D3** Hardening y resiliencia (3pt) **US-DOC1** Docs técnicas/usuario (3pt) **US-DEMO1** Presentación final (2pt)	**QA1**: Pruebas con 50 facturas, p95 API ≤5s. **D3:** Reintentos idempotentes, timeouts. **DOC1:** Guía de despliegue, runbook de incidencias, README endpoints. **DEMO1:** guión y video corto.	**R6:** Deuda técnica afecta estabilidad **M6:** Bug bash + priorización de hallazgos críticos antes de release.

Metodología Scrum (roles y ceremonias)

Roles: Product Owner (PO), Scrum Master (SM), Equipo de Desarrollo (Dev).

Ceremonias por sprint:

1. **Sprint Planning** (1 h): día 1 de cada sprint.
2. **Daily Scrum** (15 min): lun–vie.
3. **Backlog Refinement** (45–60 min): mitad de sprint.
4. **Sprint Review** (30 min): último día, con demo.
5. **Retrospective** (30 min): último día, posterior a la review.

7. PLAN DE RECURSOS

Se ha diseñado un plan que asegura que "VeriFactura" cumpla con el alcance técnico, los KPIs de aceptación y la implementación en el plazo de 6 semanas.

Recursos Humanos

- Core interno: 6–7 personas clave (~780 horas totales).
- Consultoría especializada: 4 expertos puntuales (~100 horas).
- Duración: 6 semanas.
- Enfoque: balance entre desarrollo ágil, tuning de modelos y validación bancaria.

Rol	Responsabilidades Clave	Dedicación (hrs/sem)	Total horas (6 semanas)
Project Manager (PM)	Planificación ágil (Scrum), gestión de backlog, comunicación con stakeholders, control de plazos y entregables.	20	120
Analista Funcional / de Negocio	Levantamiento de requisitos, definición de criterios de aceptación, validación con usuarios bancarios, trazabilidad de métricas.	15	90

Desarrollador Backend (API & lógica de negocio)	Implementación de módulos de recepción, filtrado, derivación de documentos, integración con OCR y LLM.	30	180
Desarrollador ML/IA (Visión + LLM)	Preprocesamiento de imágenes, integración Azure OCR, entrenamiento/evaluación de NER con LLM, generación sintética limitada.	30	180
Frontend/UI Dev (opcional/ligero)	UI mínima para visualización, edición y monitoreo de resultados.	15	90
QA/Tester	Diseño de pruebas funcionales y de precisión, ejecución UAT, validación de KPIs (F1-Score, OCR error rate, tiempos).	20	120
DevOps/Infra	Configuración de entornos (Azure, contenedores, despliegues), seguridad, monitoreo y logging.	15	90

Equipo Core interno: 780 horas en 6 semanas.

Consultoría especializada

Rol/Servicio	**Justificación**	**Estimación (horas totales)**
Consultor en Visión por Computador (Azure OCR/Pre-proc. imágenes)	Optimizar pipeline OCR, corrección de errores comunes, tuning de imágenes para facturas de baja calidad.	30 h

Consultor en LLM (NER & semántica	Ajuste fino de prompts/modelos, validación de extracción de campos críticos, retroalimentación activa.	25 h
Especialista en Seguridad Bancaria	Revisión de cifrado, anonimización de datos, cumplimiento normativo en datos sensibles (RUC, VIN).	20 h
Asesor en Datos Sintéticos	Generación controlada de facturas vehiculares sintéticas realistas para robustecer entrenamiento.	5 h

Equipo externo: 80 horas en 6 semanas.

Recursos Técnicos

Hemos estructurado estos recursos por: hardware, software y datos, considerando rol en "VeriFactura" y criticidad.

Hardware

Recurso	Descripción	Uso Principal	Observaciones
Servidores Cloud (Azure VMs)	Máquinas virtuales de 8–16 vCPUs, 32–64 GB RAM, GPU opcional (NVIDIA T4/A10).	Entrenamiento de modelos (NER, LLM fine-tuning ligero), ejecución OCR masivo, pruebas de estrés.	Escalables bajo demanda. GPU recomendada solo para entrenamiento NER.
Storage en la Nube (Blob Storage / Data Lake)	Almacenamiento seguro y estructurado de facturas (PDF, imágenes, XML).	Dataset mixto (real + sintético), logs y outputs estructurados (JSON).	Con encriptación AES-256 y control de acceso RBAC.

Base de Datos Relacional (PostgreSQL/Azure SQL)	Base para resultados estructurados (VIN, RUC, montos, etc.).	Persistencia de datos extraídos y validaciones cruzadas.	Cifrado en reposo + auditoría de queries.
Entornos de Desarrollo (Dev/QA/Prod)	Ambientes separados en la nube.	Garantizar CI/CD, pruebas y despliegues seguros.	Infra mínima con contenedores (Docker/Kubernetes).

Software

Recurso	Descripción	Uso en el Proyecto	Observaciones
Azure Vision Studio (OCR API)	OCR de alta precisión con pre-procesamiento.	Extracción de texto de PDFs/imágenes.	Pago por consumo (API calls).
Azure Cognitive Services (NER + LLM API)	Extracción semántica de entidades.	Identificación de VIN, RUC, montos, fechas, etc.	Posible ajuste con prompts o fine-tuning ligero.
Python + Frameworks (FastAPI, Flask)	Backend para pipeline de recepción y derivación.	API core de VeriFactura.	Compatible con microservicios.
Librerías IA/ML	PyTorch / TensorFlow, Hugging Face Transformers, spaCy.	Entrenamiento y validación NER, reglas post-OCR.	Uso puntual, no todo desde cero.
Pre-procesamiento Imágenes	OpenCV, Pillow.	Corrección de inclinación, ruido, contraste.	Complemento crítico para OCR.

Base de Datos	PostgreSQL/Azure SQL.	Persistencia estructurada.	Logs + auditoría para trazabilidad.
UI/Monitoreo (opcional)	React + ShadCN/UI + Tailwind.	Visualización y edición manual de facturas.	UI mínima, solo validación básica.
DevOps/Infra	Docker, Kubernetes, GitHub Actions/Azure DevOps.	CI/CD, despliegues y orquestación.	Pipelines automáticos para 3 entornos.
Seguridad	Azure Key Vault, TLS/HTTPS, JWT.	Manejo seguro de credenciales y datos sensibles.	Obligatorio por normativa bancaria.

Datos

Tipo de Dato	Fuente	Uso	Consideraciones
Facturas Vehiculares Reales (limitadas)	Scribd (colección inicial)	Entrenamiento y validación.	Acceso controlado, anonimización de datos sensibles.
Facturas Sintéticas	Generadas con scripts + plantillas + variaciones realistas.	Aumentar diversidad del dataset: errores de OCR, inclinación, distintos layouts.	Consultoría especializada valida realismo.
Ground Truth Etiquetado	Conjunto de facturas con anotaciones manuales.	Evaluación de métricas (F1-Score, OCR error rate).	Se necesita al menos 500 documentos bien etiquetados.

Logs de Procesamiento	Capturados durante pruebas/uso.	Retroalimentación para mejora continua.	Usados en retraining posterior.

Consideraciones clave en los recursos técnicos

- **Escalabilidad:** infraestructura en Azure permite ampliar CPU/GPU según demanda.
- **Seguridad Bancaria:** cifrado en tránsito (TLS 1.2+) y reposo (AES-256). Acceso con RBAC y monitoreo de auditoría.
- **Costos Cloud**: balancear entre pruebas intensivas (GPU) y producción liviana (CPU).
- **Datos Sintéticos**: deben representar escenarios de facturas reales: baja calidad de imagen, formatos mixtos, errores comunes.
- **Monitoreo KPIs**: tiempo de procesamiento, tasa de error OCR, % de intervención manual.

Recursos Financieros

- Estimado de costos por rol, consultoría externa y gastos adicionales (infraestructura, licencias, etc.).
- Los valores de referencia están ajustados según tarifas las locales corporativas:
- PM / Analista / QA: 35 USD/h
- Desarrolladores (Backend, ML/IA, Frontend): 45 USD/h
- DevOps: 40 USD/h
- Consultoría especializada externa: 80 USD/h (tarifa premium)
- Infraestructura / Licencias Azure & OCR: 2,000–3,000 USD (6 semanas, ambiente cloud controlado)

Costos del Equipo Core Interno: **USD 35,400**

Rol	Total horas (6 sem.)	Tarifa (USD/h)	Subtotal (USD)
Project Manager (PM)	120	35	4,200
Analista Funcional	90	35	3,150
Backend Developer	180	45	8,100

ML/IA Developer	180	45	8,100
Frontend/UI Dev	90	45	4,050
QA/Tester	120	35	4,200
DevOps/Infra	90	40	3,600

Costos Consultoría Externa: **USD 8,000**

Especialidad	Horas	Tarifa (USD/h)	Subtotal (USD)
Consultor en Visión por Computador (OCR)	30	80	2,400
Consultor en LLM (NER & semántica)	25	80	2,000
Especialista en Seguridad Bancaria	20	80	1,600
Asesor en Datos Sintéticos	25	80	2,000

Costos de Infraestructura y Licencias: **USD 3,200**

Concepto	Estimación (6 sem.)	Justificación
Azure OCR Vision Studio	USD 1,200	Consumo API OCR (lectura de facturas, pruebas intensivas).
Infraestructura Cloud (VMs, almacenamiento, API gateway, DBs)	USD 1,500	Ambientes Dev/QA/Prod en Azure.

Licencias / Seguridad / Logs	USD 500	Monitoreo, trazabilidad, logging seguro.

Presupuesto Total estimado:

Concepto	Monto (USD)
Equipo Core Interno	35,400
Consultoría Externa	8,000
Infraestructura & Licencias	3,200
Total Proyecto (6 semanas)	**USD 46,600**

8. HITOS Y ENTREGABLES

Hito	Fecha	Entregable	Criterios de aceptación	Responsable	Dependencias	Revisión y aprobación	Riesgos & contingencia
H1	S1-D3	Detector de texto + decisión de ruteo OCR vs parse	≥95% de acierto para detectar PDF con / sin texto en 50 PDF	AI Engineer	Dataset base	PR + demo corta Visto bueno Tech Lead	R1 OCR pobre en escaneados M1 preprocesos
H2	S1-D5	Gestor de ingesta y metadatos con cola/reintentos	Cola operativa con reintentos. ≥50 doc/min en prueba de carga mínima	Backend Engineer	H1	PR + prueba de carga Aprobación Tech Lead	R3 latencia/colas M3 batching, colas asíncronas
H3	S1-D5	Normalizador de texto	≥10% menos errores OCR vs baseline	AI Engineer	H1	PR + reporte comparativo OK Tech Lead	R1 OCR pobre en escaneados M1 preprocesos
H4	S2-D5	Integración Azure	Inferencia en lote	Backend	H2, H1	Demo con	R3 latencia/colas

		OCR en batch	con latencia < 3 seg/página	Engineer		100 págs OK Tech Lead	M3 batching, colas asíncronas
H5	S3-D3	Extracción con LLM de RUC/VIN/totales	F1 ≥0.85 en campos críticos	AI Engineer	H4	Informe F1 reproducible (F1) Revisión de datos PO	R2 variabilidad y F1 bajo M2 few-shot por concesionaria
H6	S3-D5	Validadores y reglas (VIN=17, sumas, checks)	Exactitud en sumas ≥0.98	Backend Engineer	H5	Suite de tests + reporte OK Tech Lead	R4 sesgos por concesionaria M4 estratificar validación y balancear dataset
H7	S5-D5	API REST v0 /extract + Swagger + UI	Endpoint /extract UI permite subir doc	Backend Engineer	H4-H6	Demo funcional Aprobación PO	R6 deuda técnica M6 resolver hallazgos críticos antes del release
H8	S6-D5	QA integral + Docs + Demo	50 facturas, p95 API ≤5 s	QA Engineer	H7	Ejecución plan QA Firma PO y Seguridad	R6 deuda técnica M6 resolver hallazgos críticos antes del release

* Leyenda de fecha: S = semana; D = día hábil (1–5)

9. REFERENCIAS

Katti, A. R., Reisswig, C., Guder, C., Brarda, S., Bickel, S., Höhne, J., & Faddoul, J. B. (2018). Chargrid: Towards understanding 2D documents. arXiv preprint arXiv:1809.08799. https://arxiv.org/abs/1809.08799

Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., & Jawahar, C. V. (2021). ICDAR2019 competition on scanned receipt OCR and information extraction. arXiv preprint arXiv:2103.10213. https://arxiv.org/abs/2103.10213

Park, S., Shin, S., Lee, B., Lee, J., Surh, J., Seo, M., & Lee, H. (2019). CORD: A consolidated receipt dataset for post-OCR parsing. ClovaAI. https://github.com/clovaai/cord

Bevin, V., Ananthakrishnan, P. V., Ragesh, K. R., Sanjay, M., Vineeth, S., & Wilson, B. (2025). Generating synthetic invoices via layout-preserving content replacement. arXiv preprint arXiv:2508.03754. https://arxiv.org/abs/2508.03754

Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., & Park, S. (2021). OCR-free document understanding transformer. arXiv preprint arXiv:2111.15664. https://arxiv.org/abs/2111.15664

Lin, L. C. (2025, abril 24). From RPA to GenAI: How Uber's invoice breakthrough points the way. Lewis C. Lin. https://www.lewis-lin.com/blog/from-rpa-to-genai-how-ubers-invoice-breakthrough-points-the-way

Scribd. (2025). Factura vehículo. Scribd. https://www.scribd.com/document/440281304/Factura-vehiculo

Šimsa, Š., Šulc, M., Uřičář, M., Patel, Y., Hamdi, A., Kocián, M., Skalický, M., Matas, J., Doucet, A., Coustaty, M., & Karatzas, D. (2023). DocILE: Benchmark for document information localization and extraction. arXiv preprint arXiv:2302.05658 / Rossum.ai. https://docile.rossum.ai/

Retnan, R. (2025, julio 7). Automating invoice data extraction: OCR vs LLMs explained. RaftLabs (Medium). https://raftlabs.medium.com/automating-invoice-data-extraction-ocr-vs-llms-explained-f75fc1596ef0

Sánchez, J., Salgado, A., García, A., & Monzón, N. (2022). IDSEM, an invoices database of the Spanish electricity market. Scientific Data, 9(786). https://doi.org/10.1038/s41597-022-01885-3